Fundamental Limitations of Alignment in Large Language Models

Year: 2023 | *Citations: 168* | *Authors: Yotam Wolf, Noam Wies, Yoav Levine*

Abstract

An important aspect in developing language models that interact with humans is aligning their behavior to be useful and unharmful for their human users. This is usually achieved by tuning the model in a way that enhances desired behaviors and inhibits undesired ones, a process referred to as alignment. In this paper, we propose a theoretical approach called Behavior Expectation Bounds (BEB) which allows us to formally investigate several inherent characteristics and limitations of alignment in large language models. Importantly, we prove that within the limits of this framework, for any behavior that has a finite probability of being exhibited by the model, there exist prompts that can trigger the model into outputting this behavior, with probability that increases with the length of the prompt. This implies that any alignment process that attenuates an undesired behavior but does not remove it altogether, is not safe against adversarial prompting attacks. Furthermore, our framework hints at the mechanism by which leading alignment approaches such as reinforcement learning from human feedback make the LLM prone to being prompted into the undesired behaviors. This theoretical result is being experimentally demonstrated in large scale by the so called contemporary"chatGPT jailbreaks", where adversarial users trick the LLM into breaking its alignment guardrails by triggering it into acting as a malicious persona. Our results expose fundamental limitations in alignment of LLMs and bring to the forefront the need to devise reliable mechanisms for ensuring AI safety.